Proxy Variables and the Racial Gap in Voided Ballots

Michael Tomz
Stanford University[1]

Robert P. Van Houweling
University of Michigan[2]

December 16, 2002

[1]Assistant Professor, Department of Political Science, Encina Hall West, Stanford, CA 94305-6044 (tomz@stanford.edu).

[2]Assistant Professor of Political Science, Gerald R. Ford School of Public Policy, 440 Lorch Hall, University of Michigan, Ann Arbor, MI 48109-1220 (rpvh@umich.edu). Part of this research was conducted while Van Houweling was a visiting scholar at Stanford University.

Abstract

Nearly all work on the connection between race and voided ballots relies on proxy variables, such as the nonwhite proportion of either registered voters or the general population, to characterize those who actually went to the polls. This research note discusses the consequences of using proxy variables. When the assumptions for unbiased ecological regression are otherwise satisfied, the use of proxies introduces nonrandom measurement error that depresses estimates of African American and white invalidation, as well as the difference between the two rates. A recent paper by Tomz and Van Houweling (2003) avoids these problems by employing direct measures of racial turnout.

1 The Use of Proxy Variables

Nearly all work on the connection between race and voided ballots relies on proxy variables, such as the nonwhite proportion of either registered voters or the general population, to characterize those who actually went to the polls. This research note discusses the consequences of using proxy variables. When the assumptions for unbiased ecological regression are otherwise satisfied, the use of proxies introduces nonrandom measurement error that depresses estimates of African American and white invalidation, as well as the difference between the two rates.

Using the notation in Tomz and Van Houweling (2003), the fraction of invalid ballots in any particular precinct is given by the accounting identity $I = \beta_n T_n + \beta_w(1 - T_n)$. With proxy variables, the identity can be written as $I = \gamma_n R_n + \gamma_w(1 - R_n)$, where R_n and $(1 - R_n)$ represent the nonwhite and white proportions of the registered electorate, as distinct from voters who actually showed up at the polls. This approach causes no bias in the special case in which whites and minorities turn out at identical rates. To see this, note that $T_n = \frac{\tau_n R_n}{\tau_n R_n + \tau_w(1-R_n)}$, where τ_n and τ_w are the nonwhite and white rates of turnout, respectively. If $\tau_n = \tau_w$, this equation reduces to $T_n = R_n$, meaning that minority turnout as a proportion of total turnout exactly mirrors the minority share of the registered population. Under these unusual conditions, the use of the proxy will not affect the conclusions.

Once we acknowledge that turnout varies by race, though, a complicated form of measurement error emerges. The decision to use registration in place of turnout amounts to replacing the true explanatory variable, T_n, with the proxy αT_n, where α is a non-linear scale

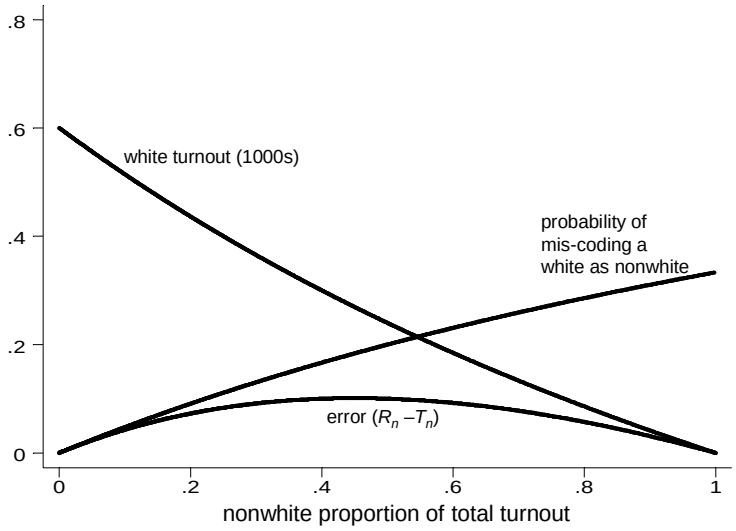
factor $\frac{\tau_w}{T_n(\tau_w - \tau_n) + \tau_n}$. The greater the difference between τ_n and τ_w, the more this scale factor will deviate from unity. Note that α does not affect all observations equally. The impact on any particular observation depends on the nonwhite share of total turnout (T_n), which varies across precincts depending on their racial composition. If whites turn out at a higher rate than minorities, α will be larger in heavily white districts than in minority-dominated ones. Clearly, the use of R_n introduces non-random measurement error into the explanatory variable, thereby contaminating estimates of racial invalidation rates.

2 The Potential for Bias with Proxy Variables

The net effect on the regression line will depend on the frequency of white versus nonwhite precincts. Suppose that minorities invalidate their ballots at a higher rate than whites (Tomz and Van Houweling 2003). If most precincts have white majorities, as in the United States, then the researcher who regresses I on R_n and a constant term will underestimate the invalidation rates of whites and nonwhites, and also understate the difference between the two. If, contrary to fact, most precincts had white minorities, then the procedure would still underestimate nonwhite and white spoilage, but it would overstate the racial gap.

The reasons are complicated but important to consider. Since τ_w is typically greater than τ_n, we know that $\alpha > 1$, which implies that the proxy exaggerates the nonwhite share of total turnout. Using the proxy is, therefore, tantamount to misclassifying whites as nonwhites. The probability of mistaking any particular white as a nonwhite, $(R_n - T_n)/(1 - T_n)$, increases as the precinct becomes less white, but the raw number of whites who could potentially be

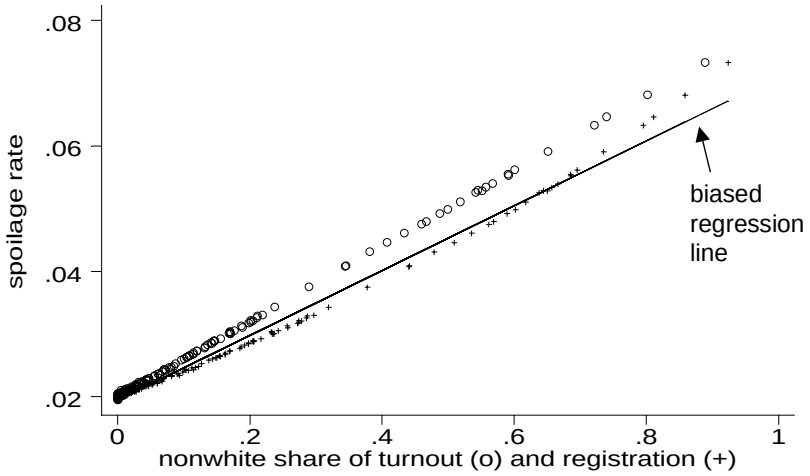
Figure 1: Measurement Error Due to Proxy Variables



mis-identified falls in tandem. Figure 1 illustrates these opposing effects. The figure is based on the assumption that each precinct has 1000 registered voters, that whites turn out at a rate of 60% and that nonwhites turn out at a rate of 40%. The figure shows that the net effect on the proxy variable is parabolic: measurement error $(R_n - T_n)$ is greatest in precincts with a rough balance of whites and nonwhites, where the discrepancy between the proxy and the true explanatory variable is approximately 10 percentage points.

This measurement error reduces the estimated invalidation rate for both minorities and whites. The first effect is rather intuitive: if ballots of white voters go uncounted at a lower rate than those of minorities, and if registration data systematically misrepresent whites as minorities, then the procedure attributes to minorities some of the low-spoilage behavior by whites. Figure 2 shows this phenomenon. The figure displays the precinct-wide invalidation rate (on the vertical axis) against both T_n (represented by the o's) and R_n (represented by

Figure 2: How Proxy Variables Introduce Bias



the +'s), under the assumption that $\beta_n = .08$ and $\beta_w = .02$.[1] Note that every o has a corresponding +, but the +'s are shifted anywhere between 0 and 10 percentage points to the right, reflecting the parabolic measurement error that was depicted in Figure 1. Thus, if we base the analysis on the +'s rather than the o's, it takes a higher proportion of nonwhites to achieve any given level of invalidation. Put another way, using registration data shifts the entire regression line downward and to the right. As a result, the analyst will underestimate β_n.

The proxy-based approach also depresses the estimate of invalid ballots cast by whites. This may come as a surprise to readers, since the aggregate number of uncounted ballots has not changed. If one group is estimated to cast invalid ballots at a lower rate, must

[1] We also assume that nonwhites make-up approximately 12% of the registered electorate nationwide, though we allow the proportion in any particular precinct to take on any value. To satisfy these conditions, we randomly generated the racial data from a beta distribution with parameters $\alpha = 0.25$ and $\beta = 1.75$.

not the other group cast invalid ballots at a higher rate to account for the difference? This intuition, though appealing, is incorrect. Recall that the standard procedure "removes" whites from all precincts but subtracts them at the highest rate in heavily nonwhite areas, which tend to have the largest proportion of lost votes. As a result, whites appear to be clustered in low-invalidation precincts, reducing their average invalidation rate as well. Thus, the proxy variable deflates the estimate of invalidation rates by nonwhites, but it also raises the apparent number of nonwhite voters, allowing minorities to account for a higher proportion of the voided ballots. The effective increase in nonwhite voters more than compensates for their reduced rate of invalidation, leading to a lower estimate of β_w. The sharper the difference in turnout between whites and minorities, the more negative the bias becomes. Using the racial breakdown of the total population, an even greater distortion of racial turnout (but one that is necessary in forty-three US states that do not report voter registration by race), would only exacerbate the bias.

Finally, researchers who use R_n in place of T_n will tend to underestimate the racial gap in voided ballots. Why is this the case? In a country where the median precinct has a white majority, using the proxy variable not only shifts but also flattens the regression line, due to the large number of predominantly white precincts. Figure 2 displays this phenomenon: the dense +'s in the lower left quandrant exert a strong effect on the regression line, which the sparse +'s in the upper right quadrant cannot overcome. As a consequence, the estimate of $\beta_n - \beta_w$ will shrink. We illustrate this effect in Figure 3, which plots the gap that one might estimate with R_n against the gap using the correct explanatory variable, T_n. The

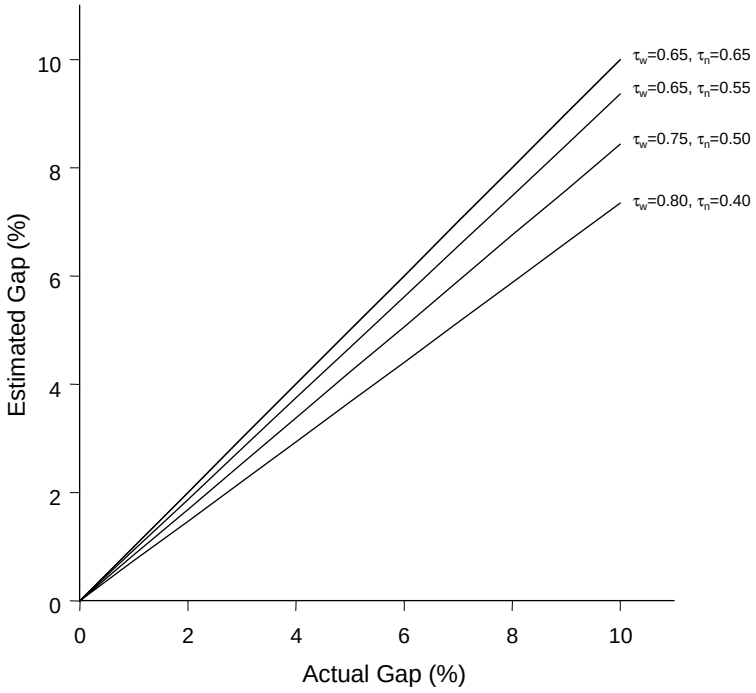
45-degree line emanating from the origin marks the set of coordinates in which the estimated gap exactly matches the actual one. As discussed earlier, this perfect correspondence will arise only when both racial groups turn-out at equal rates, e.g., $\tau_w = \tau_n = 0.65$ in this figure. When turnout rates differ, the actual gap will exceed the estimated one. The bias, represented by the vertical distance between the 45-degree line and the other rays, increases with the true disparity in both invalidation and turnout rates.

3 Conclusion

To conclude, researchers who use proxy variables in a standard Goodman's regression will typically underestimate β_n, β_w, and the difference between them. Tomz and Van Houweling (2003) avoid this bias by focusing on data from Louisiana and South Carolina, the only U.S. states that officially report turnout by race.[2]

[2]It may also be possible to address the bias with statistical procedures. One option, known as double-regression, is a "statistical trick ... to cope with the nonlinearity resulting from differential turnout by race" (Grofman 1993: 482; see also King 1997: 71-73). The procedure would involve running two separate regressions, one to explain turnout and the other to explain spoilage, and then combining the results to obtain an estimate of the spoilage rate for each race. As an alternative, researchers might introduce a quadratic term, e.g., $E(I) = b_1 + b_2 R_b + b_3 R_b^2$, to account for the arc-shaped measurement error. Figure 2 shows that the use of R_n leads to a non-linear relationship between invalidation and the racial composition of the district. Technically, this phenomenon is known as aggregation bias, which arises in this case because the parameter of interest (the invalidation rate) is correlated with the regressors. The proxy-based model can be written as $E(I) = \gamma_n \alpha T_n + \gamma_w \alpha (1 - T_n)$, where $\gamma_n \alpha$ and $\gamma_w \alpha$ are supposed to represent the invalidation rates. Recall that α varies with T_n, which implies that the invalidation rates in this specification are functions of the regressors T_n and $(1 - T_n)$. Seen in this way, it becomes clear that the proxy-based approach introduces aggregation bias into the model. A quadratic term could minimize the aggregation bias. To our knowledge, previous work on racial patterns of voided ballots has not taken advantage of either technique.

Figure 3: Proxy Variables Shrink the Estimated Gap in Voided Ballots



References

Grofman, Bernard. 1993. "Throwing Darts at Double Regression – and Missing the Target." Social Science Quarterly 74, no. 3 (September): 480-87.

King, Gary. 1997. A Solution to the Ecological Inference Problem. Princeton, NJ: Princeton University Press.

Tomz, Michael and Robert P. Van Houweling. 2003. "How Does Voting Equipment Affect the Racial Gap in Voided Ballots?" American Journal of Political Science 47, no. 1 (January 2003): 45-59.